

A stochastic model of evolution

Hervé Guiol, Fábio P. Machado and Rinaldo B. Schinazi

TIMB-TIMC Univ. Grenoble, France, IME-USP, Brasil, Math. Dept. UCCS, USA

August 24th 2009

Abstract. We propose an alternative to the Bak-Sneppen model for species survival. In our model the number of species is random. Births and deaths occur with constant probabilities and it is only the species with lowest fitness which is removed. We show that there is a sharp phase transition when the birth probability is larger than the death probability. The set of species with fitness higher than a certain critical value approach an uniform distribution. On the other hand the species with fitness less than the critical disappear after a finite (random) time.

1. Introduction. In the Bak-Sneppen model, see Bak and Sneppen (1993), there is a fixed number N of species arranged in a circular graph. At each discrete time the site on the circle with the lowest fitness and its two nearest neighbors have their fitness replaced by a random number independently sampled from the uniform distribution on $[0, 1]$. There are a number of exciting results and conjectures for the Bak-Sneppen model, see Gillett, Meester and van der Wal (2006). However, from a biological point of view the model does seem a little artificial. We propose a model that seems more natural to us. In our model the number of species is random (instead of a fixed number) and we replace only the species with the lowest fitness (instead of replacing its nearest neighbors as well).

Note that replacing the species with the lowest fitness involves a global interaction of all the species present. The Bak-Sneppen model involves this global interaction and a local interaction (replacing the neighbors). As pointed out by Meester and Znamenski (2003) if only the lowest fitness were replaced then the Bak-Sneppen system would converge to a configuration with all fitnesses equal to 1 which makes the model not so interesting. Even though our model involves only the global interaction it has a non trivial limiting behavior. We now give the details of our model.

Consider a discrete time model that starts from the empty set. At each time $n \geq 1$ with probability p there is a birth of a new species and with probability $q = 1 - p$ there is a death of a species (if the system is not already empty). Hence, the total number

Key words and phrases: Bak-Sneppen model, species survival, stochastic model

of species at time n is a random walk on the positive integers which jumps to the right with probability p and to the left with probability q . When the random walk is at 0 then it jumps to 1 with probability p or stays at 0 with probability $1 - p$. Each new species is associated with a random number. This random number is sampled from the uniform distribution on $[0, 1]$. We think of the random number associated with a given species as being the fitness of the species. These random numbers are independent of each other and of everything else. Every time there is a death event then the type that is killed is the one with the smallest fitness. This is similar to a model introduced by Liggett and Schinazi (2009) for a different question.

Take p in $(1/2, 1)$ and let

$$f_c = \frac{1 - p}{p}.$$

Note that f_c is in $(0, 1)$. Let L_n and R_n be the set of species alive at time n whose fitness is lower and higher than f_c , respectively. Since each fitness appears at most once we can identify each species to its fitness and think of L_n and R_n as sets of points in $(0, f_c)$ and $(f_c, 1)$, respectively. Let R'_n be the set of species that have appeared up to time n (they can be dead or alive at time n) and whose fitness is higher than f_c . Clearly, at all times n , R_n is a subset of R'_n . Let $|A|$ denote the cardinal of set A . We are now ready to state our main result.

Theorem. *For every $\epsilon > 0$ we have almost surely for n large enough that*

$$0 \leq |R'_n| - |R_n| \leq \frac{2}{pf_c} n^{1/2+\epsilon}.$$

In particular, the set R_n of species whose fitness is above f_c approaches a set of i.i.d. uniform random variables on $(f_c, 1)$.

Observe that the number of species in R'_n is a random variable that follows a binomial distribution with parameters n and $p(1 - f_c)$. Moreover, R'_n is a (random) set of i.i.d. uniform random variables on $(f_c, 1)$. Hence, the set R_n of species whose fitness is higher than f_c approaches an uniform system because the number of discrepancies between R_n and R'_n gets smaller and smaller compared to the number of species (which is of order n at time n). Interestingly, the same type of uniform behavior in some $(f_c, 1)$ is expected for the Bak-Sneppen model but this is still unproved, see Meester and Znamenski (2003).

We now discuss the behavior of L_n (the set of species whose fitness is lower than f_c). Observe that $|L_n|$ (the cardinal of L_n) increases by 1 with probability pf_c , decreases by 1 with probability q (if it is not already at 0) and stays put with probability $p(1 - f_c)$. Since $pf_c = q$, $|L_n|$ is symmetric and it is easy to check that it is null recurrent. That

is, $|L_n|$ will hit 0 infinitely many times but the expected time of return to 0 is infinite. So there will be times when $|L_n|$ is large but it will always eventually come back to 0. Because we kill the species with lowest fitness the surviving species will tend to be concentrated close to f_c . See figure 1.

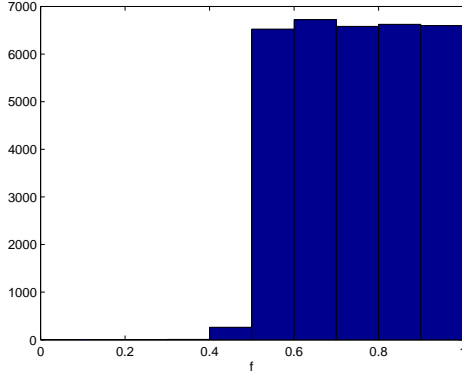


Figure 1. This is the histogram of the fitnesses after 100,000 births and deaths for $p = 2/3$. We have $f_c = 1/2$ and as predicted by the Theorem the distribution on $(f_c, 1)$ approaches an uniform.

Observe that the larger p is the more welcoming the environment is to new species. If p is only slightly larger than $1/2$ then f_c is close to 1 and only species with high fitness will survive. On the other hand if p is close to 1 then f_c is close to 0 and even species with relatively low fitness will survive.

2. Proof of the Theorem.

The idea of the proof is quite simple. For times n for which L_n (the set of species whose fitness is lower than f_c) is not empty all deaths must occur in L_n (and no species in R_n can die). Hence, at these times n , R_n and R'_n are coupled. Let t_n be the number times $k \leq n$ for which L_k is empty. That is,

$$t_n = |\{1 \leq k \leq n : L_k = \emptyset\}|.$$

We will show that, for any $\epsilon > 0$, t_n is almost surely less $n^{1/2+\epsilon}$ for n large enough. Hence, the number of deaths in R_n is at most $n^{1/2+\epsilon}$. The main step in the proof is the following Lemma.

Lemma. *There are positive constants γ and D such that for every $\epsilon > 0$ we have*

$$P\left(t_n > \frac{2}{pf_c} n^{1/2+\epsilon}\right) \leq D \exp(-\gamma n^\epsilon).$$

Once the Lemma is proved we are done. For we use the Borel-Cantelli Lemma to show that almost surely there is a natural N such that $t_n \leq Cn^{1/2+\epsilon}$ for $n \geq N$, where $C = \frac{2}{pf_c}$. This shows that the number of discrepancies between R_n and R'_n is at most $Cn^{1/2+\epsilon}$.

Proof of the Lemma.

Recall that we start from the empty set. After a geometric random time with mean $\frac{1}{pf_c}$, denoted by G_0 , the first species appears in $(0, f_c)$. That is,

$$G_0 = \min\{k \geq 1 : L_k \neq \emptyset\}.$$

Let

$$E_1 = \min\{k \geq G_0 : L_k = \emptyset\}.$$

Hence, E_1 be the time it takes starting at time G_0 for $|L_\cdot|$ to return to 0. More generally, we define for $i \geq 1$

$$G_i = \min\{k \geq G_0 + E_1 + \dots + G_{i-1} + E_i : L_k \neq \emptyset\},$$

and

$$E_{i+1} = \min\{k \geq G_0 + E_1 + \dots + E_i + G_i : L_k = \emptyset\}.$$

Note that the $(G_i)_{i \geq 0}$ and the $(E_i)_{i \geq 1}$ are two i.i.d. sequences. Moreover, the G_i follow a geometric distribution with mean $\frac{1}{pf_c}$.

Let k_n be the number of times that L_k hits the empty set by time n :

$$k_n = |\{2 \leq k \leq n : |L_{k-1}| = 1 \text{ and } |L_k| = 0\}|.$$

That is, k_n counts the number of times L_k goes from 1 to 0 species for $k \leq n$. Note that if $k_n = 0$ then $t_n \leq G_0$. We now compute

$$(1) \quad P(t_n > Cn^{1/2+\epsilon}) \leq P(t_n > Cn^{1/2+\epsilon}; k_n < n^{1/2+\epsilon}) + P(k_n \geq n^{1/2+\epsilon}).$$

For $k_n \geq 1$ we have

$$G_0 + G_1 + \dots + G_{k_n-1} < t_n \leq G_0 + G_1 + \dots + G_{k_n},$$

and for $k_n = 0$ we have $t_n \leq G_0$. Hence,

$$P(t_n > Cn^{1/2+\epsilon}; k_n < n^{1/2+\epsilon}) \leq P(G_0 + G_1 + \dots + G_{m_n} > Cn^{1/2+\epsilon}),$$

where m_n is the integer part of $n^{1/2+\epsilon}$. Now, the expected value of $G_0 + G_1 + \dots + G_{m_n}$ is $\frac{m_n+1}{pf_c}$. By a standard large deviations inequality (see for instance Lemma (9.4) in Chapter 1 of Durrett (1996)) there exists $\gamma > 0$ such that

$$(2) \quad P(G_0 + G_1 + \dots + G_{m_n} > Cn^{1/2+\epsilon}) \leq \exp(-\gamma m_n) \leq \exp(-\gamma(n^{1/2+\epsilon} - 1)).$$

We now take care of the second term in the r.h.s. of (1). Observe that $|L_n|$ increases by 1 with probability pf_c , decreases by 1 with probability q (if it is not already at 0) and stays put with probability $p(1-f_c)$. Since $pf_c = q$, $|L_n|$ is essentially a simple symmetric random walk with a reflecting barrier at 0. In particular it is a recurrent Markov chain and it will come back to 0 infinitely often.

Now using that the E_i are i.i.d. and that for $1 \leq i \leq k_n - 1$ they all must be less than n ,

$$P(k_n \geq n^{1/2+\epsilon}) \leq P(E_1 < n)^{m_n-1}.$$

It is easy to couple $|L_n|$ to a genuine simple symmetric random walk (one that jumps +1 or -1 with probability 1/2 at each step) so that $|L_n|$ takes more time to go from 1 to 0 than the genuine walk does. Let T_0 be the time for a genuine simple symmetric random walk to hit 0. We have

$$P(k_n \geq n^{1/2+\epsilon}) \leq P(E_1 < n)^{m_n-1} \leq P_1(T_0 < n)^{m_n-1}.$$

It is well known that $P_1(T_0 \geq n)$ is asymptotically $1/\sqrt{\pi n/2}$, see for instance Chapter III in Feller (1968). Hence, there are constants $\gamma' > 0$ and D such that

$$(3) \quad P(k_n \geq n^{1/2+\epsilon}) \leq \exp(-\gamma' \frac{m_n-1}{n^{1/2}}) \leq D \exp(-\gamma' n^\epsilon).$$

Using (2) and (3) in (1) completes the proof of the Lemma and therefore of the Theorem.

Acknowledgements H.G. thanks IXXI for partial support, Math. Department of University of Colorado at Colorado Springs and IME USP Brasil for kind hospitality. F.M. thanks Math. Department of University of Colorado at Colorado Springs for partial support and kind hospitality. R.B.S. was partially supported by N.S.F. grant DMS-0701396.

References.

P. Bak and K. Sneppen (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, **74**, 4083-4086.

R. Durrett (1996). *Probability: theory and examples* (second edition). Duxbury Press.

W. Feller (1968). *An introduction to probability theory and its applications*. John Wiley (3rd edition).

A. Gillett, R. Meester and P. van der Wal (2006). Maximal avalanches in the Bak-Sneppen model. *J. Appl. Prob.*, **43**, 840-851.

T. Liggett and R. B. Schinazi (2009). A stochastic model for phylogenetic trees. To appear in the *J. Appl. Prob.*

R. Meester and D. Znamenski (2003). Limit behavior of the Bak-Sneppen evolution model. *Ann. Prob.*, *31*, 1986-2002.